

INVESTIGATING SPAM MASS VARIATIONS FOR DETECTING WEB SPAM

Muhammad Atif Quresh, Arjumand Younus, Nasir Touheed

Institute of Business Administration

e-mail: maqureshi@iba.edu.pk, ayounus@iba.edu.pk, ntouheed@iba.edu.pk

Abstract

In this paper, we investigate variations of Spam Mass for filtering web spam. Firstly, we propose two strategies for designing new variations of the Spam Mass algorithm. Then, we perform experiments among different versions of Spam Mass using WEBSPAM-UK2006 data set. Finally, we show improvement through proposed strategy by up to 1.33 times in recall and 1.02 times in precision over the original version of Spam Mass.

Keywords: *Web algorithms, Web spam filtering, Spam Mass variations, link spam, link analysis.*

1. INTRODUCTION

Web spam is an activity of deceiving the ranking function of a web search engine by employing techniques to boost the importance of a web page without improving the quality in claimed service [1]. Since the birth of popular PageRank algorithm [2] web search engines have opted to exploit web link structure for calculating importance of a web page [7, 11]. Spam Mass [5] is a popular web spam detection algorithm which exploits the web link structure for the discovery of web spam. In this paper, we investigate the winner (in terms of quality improvement) among the different versions of Spam Mass: original [5], based upon modified TrustRank [9] and our newly proposed variations.

The remainder of this paper is organized as follows. In Section 2 we present web graph model, PageRank, TrustRank and Anti-TrustRank. In Section 3 we present Spam Mass and a known variation of TrustRank, Anti-TrustRank and Spam Mass [9]. In Section 4 we introduce our proposed strategies for designing new variations of Spam Mass. In Section 5 we show the results of our experimentation. Finally we conclude our paper in Section 6.

2. BACKGROUND

In this section, we show how the web can be modeled as a directed graph and after that we show PageRank algorithm that changed the way how ranking is done in modern search engines [4, 6]. Then, we present TrustRank algorithm which was an attempt to patch shortcoming of PageRank (by punishing web spam) [6]. Finally, we present Anti-TrustRank algorithm which outperformed TrustRank in punishing web spam [8].

2.1. Web Graph Model

Web can be modeled as a directed $G=\{V, E\}$, where V represents web nodes (simply, vertices) and E represents directed links (simply, edges). Incoming links to a web node are known as *inlinks* and outgoing links from a web node are known as *outlinks*. A web node may be defined at any level of abstraction such as web page, domain, sub-domain, host, etc.

2.2. PageRank

PageRank is a popular ranking algorithm which exploits the web link information for assigning global importance to the web nodes. It works on the philosophy that a web node is more important if it is inlinked by many unimportant web nodes or by at least some important web nodes. The PageRank score of a web node is defined as follows.

$$PR[p] = d \cdot \sum_{q:(q,p) \in E} \frac{PR[q]}{N_{outlink}(q)} + (1 - d) \cdot v[p] \quad (i)$$

Where $PR[p]$ represents the PageRank score of the web node p , d is the probability of following an outlink, (q, p) is the set of those web nodes which outlinks to the web node p , $N_{outlink}(q)$ is the number of outlinks from q , $v[p]$ is the probability (which is uniform) that the user does not follow an outlink instead he or she takes a random jump to the web node p (e.g., by typing a URL in the address bar).

PageRank's estimation for assigning global importance to web nodes can be exploited by creating a bogus web link structure i.e., outlinking heavily to an unimportant web node. This is analogous to fake votes in a voting process.

2.3. TrustRank

TrustRank is a variant of PageRank which exploits the importance of trust in the web link structure for assigning global importance to the web nodes. The philosophical difference between PageRank and TrustRank is that every vote in a voting process is not equal in weight. Thereby, TrustRank employs the factor of trust for biasing the worth of a vote (which is received in the form of an inlink from a trusted web node). TrustRank begins by taking an input seed set of trusted web nodes (such as *.gov, *.edu). Then, it propagates scores from the input seed set of web nodes to their outlinks in an iterative way and generates a trust score vector of entire web nodes. After that, it estimates global importance of web nodes by biasing ranking function with trust scores. In addition, TrustRank assigns the value of random jump of a web node based upon trust score of that web node.

2.4 Anti-TrustRank

Anti-TrustRank is philosophically opposite to that of how TrustRank calculates trust scores of web nodes. Precisely, Anti-TrustRank takes an input seed set of web nodes which are involved in web spam or have a bad reputation. Then, it propagates scores in reverse direction (i.e., inlinks) in an iterative way and generates anti-trust scores. Web nodes that have high anti-trust scores are declared as spam nodes (i.e., involved in web spam). In [8], Anti-TrustRank was shown to outperform TrustRank. However, Anti-TrustRank and TrustRank take different input seed sets and therefore, a comparison between them would not be fair a comparison.

3. RELATED WORK

In this section, we explain original Spam Mass which uses TrustRank for detecting web spam [GBG05]. After that we explain a variation of each algorithm: TrustRank, Anti-TrustRank and Spam Mass which outperforms their respective original version [9].

3.1. Spam Mass

Spam Mass is a web spam detection algorithm which bases itself on top of PageRank and TrustRank. In [9], Spam Mass is found to be more effective than Anti-TrustRank. The basic idea is that PageRank represents overall score received due to web spam and normal activity, and TrustRank is representative of normal activity, therefore, if a PageRank score of a web node is much higher than its TrustRank score then such a web node is involved in web spam. Following relation shows the constraint for declaring a web node as spam node.

$$\frac{PR[p] - TR[p]}{PR[p]} \geq \tau \quad (ii)$$

Here $PR[p]$ represents PageRank score of web node p , $TR[p]$ represents TrustRank score of web node p and τ (called relativeMass) represents the threshold value across which comparison is made. It is important to note that PageRank scores and TrustRank scores are normalized so that the scale remains same.

3.2. Double Seeded TrustRank

In [9], double seeded TrustRank (simply, modified version [9]) is found to be better than respective original version. The basic idea of this variation is to stop propagation of trust scores to the outlinks pointing to a spam node and for this an additional seed set of spam nodes is provided.

A good example when a trusted node may outlink to a spam node can be when a university student outlinks to his friend who is involved in web spam. In this variation, trust score is only propagated to the outlinks which are not pointing to the seed set of spam nodes.

3.3. Double Seeded Anti-TrustRank

In [9], double seeded Anti-TrustRank (simply, modified version [9]) is found to be better than respective original version. Double seeded Anti-TrustRank is based on similar basic idea as defined in Section 3.2 i.e., a web spam node may be inlinked by a trusted web node, therefore, for the sake of prevention, an additional seed set of trusted web nodes is provided at input. In this algorithm, anti-trust score is restricted to propagate to only those inlinks which are not pointing to the seed set of trusted web nodes.

3.4. Double Seeded Spam Mass

In [9], double seeded Spam Mass (simply, modified version [9]) is found to be better than respective original version. The basic idea is similar to that presented in Section 3.2 and therefore the relation presented in (ii) is changed i.e., replacing TrustRank scores with double seeded TrustRank scores.

4. PROPOSED VARIATIONS

In this section, we explain our two proposed strategies for designing new variations of Spam Mass algorithm.

4.1. Reflected Double Seeded Spam Mass

In the basic philosophy of Spam Mass, TrustRank is used as a measure to show rank contribution coming from trusted web nodes and PageRank is used as a measure to show overall rank contribution. In this strategy, we propose different ranking algorithms in place of PageRank and variation of TrustRank as discussed in Section 3.2. For example, we may use variation of Anti-TrustRank (mentioned in Section 3.3) instead of the variation of TrustRank by making slight adjustment. Philosophically, anti-trust score shall be used as a representative of spam nodes instead of trusted web nodes, therefore, we took a philosophical dual of anti-trust scores. After taking that, highest anti-trust score gets converted into lowest score and lowest anti-trust score becomes highest score. This makes sense because a web node with highest anti-trust score should have lowest trust score and vice versa. Dual of anti-trust scores may be calculated as follows.

1. $max = Max(ATR)$.
2. $min = Min(ATR)$.
3. For Each $n \in N$: $DualATR[n] = max - ATR[n] + min$.

Figure 1. Steps for calculating Dual

Here, N is the set of entire web nodes.

Following relation shows the constraint of reflected double seeded Spam Mass by replacing variation of TrustRank (as in Section 3.2) with the Dual of variation of Anti-TrustRank (as in Section 3.3).

$$\frac{PR[n] - Dual(ATR_ds[n])}{PR[n]} \geq \tau \quad (iii)$$

Here, Dual function is the dual as explained in Fig. 1 and ATR_ds represent double seeded Anti-TrustRank score of web node n .

Similarly, other possible constraints could be made by replacing PageRank with a variation of Anti-TrustRank while keeping the Dual of variation of Anti-TrustRank (as in equation iii) or variation of TrustRank (as in Section 3.2), etc.

4.2. Inclusive Double Seeded Spam Mass

In this strategy, we propose to use overall rank contribution, rank contribution coming from trusted web nodes, and rank contribution coming from spam web nodes for defining the constraint. One of the possible constraints for declaring a web node as a spam node may be shown as follows.

$$\frac{PR[n] - (TR_ds[n] - ATR_ds[n])}{PR[n]} \geq \tau \quad (iv)$$

Where, TR_ds represents double seeded TrustRank score of web node n .

We can produce other similar constraints by involving dual of variation of TrustRank and Anti-TrustRank.

5. EXPERIMENTAL RESULTS

In this section, we present the results of our evaluation among the versions of Spam Mass algorithm.

5.1. Dataset

We use WEBSpAM-UK2006 data set [3] which is a collection 77.9 Million web pages of 11,402 hosts. This dataset is partially labeled in terms of trusted (non-spam) and spam hosts (i.e., spam nodes). Furthermore, the data set is classified into two sets ‘input seed set’ and ‘test set’. Input seed set is divided into non-spam seed set (4,948 hosts) and spam seed set (674 hosts). Similarly, test set is also divided into non-spam set (601 hosts) and spam set (1,250 hosts).

Input seed set serves as an input for the web spam detection algorithm while the output from the algorithm is compared across the test set.

We perform all experiments by using web graph with hosts being the web nodes.

5.2. Measures and Notations

We use precision and recall [10] as the measures for performing evaluation among the versions of Spam Mass algorithm. Table 1 shows the notations used for explaining experiments.

Table 1. Notations

Notations	Description
<i>Pre</i>	Precision
<i>Re</i>	Recall
<i>Top</i>	Web nodes among the top percentage of PageRank scores
<i>SM</i>	Original Spam Mass algorithm
<i>SM_ds</i>	Double seeded Spam Mass algorithm
<i>SM_rds</i>	Reflected double seeded Spam Mass
<i>SM_rds_pr_dualmatr</i>	Reflected double seeded Spam Mass algorithm with following relation $\frac{PR[n] - Dual(ATR_ds[n])}{PR[n]} \geq \tau$
<i>SM_rds_matr_dualmatr</i>	Reflected double seeded Spam Mass algorithm with following relation $\frac{ATR_ds[n] - Dual(ATR_ds[n])}{ATR_ds[n]} \geq \tau$
<i>SM_rds_matr_mtr</i>	Reflected double seeded Spam Mass algorithm with following relation $\frac{ATR_ds[n] - TR_ds[n]}{ATR_ds[n]} \geq \tau$
<i>SM_ids_pr_mtr_matr</i>	Inclusive double seeded Spam Mass algorithm with following relation $\frac{PR[n] - (TR_ds[n] - ATR_ds[n])}{PR[n]} \geq \tau$

5.3. Experiments

Before performing our evaluation we apply well-known labeling rule [1, 3, 5, 6, 8] which is a general practice for marking hosts as non-spam and spam. Hosts pertaining to education and government [1, 3, 5, 6] are labeled as non-spam (i.e., *.ac.uk, *.gov.uk, *.police.uk). Similarly, hosts

pertaining to spam terms in their hostnames (terms such as mp3, mortgage and sex [1,8]) are labeled as spam.

After expanding input seed set with the above mentioned labeling rule we conduct experiments for evaluation. In original paper of Spam Mass [5], authors pointed out that finding web spam from entire data set is not important, and we believe that it is due to the reason that users are interested in the top results of a search engine instead of entire match. Therefore, keeping top results in mind and completeness of web spam detection, we perform experiments i.e., detecting web spam among the top 10%, 50% and 100% (entire) of PageRank scores.

Results of web spam detection

We pick six important possible variations of Spam Mass algorithm (due to space limitation) for showing the comparison. Figs. 2–4 show the results of experiments among top 10%, 50%, and 100% of PageRank scores respectively. In each figure we show precision and recall when *relativeMass* is varied from 0.95–1.0. In original Spam Mass [5] authors stated 0.98 to be a reasonable value for applying Spam Mass and in [9], 0.99 was found to be reasonable value. Moreover, for the sake of readability we show Table 2 which is a snapshot of readings when *relativeMass* is varied between 0.98–1.0.

Figs. 2–4 and Table 2 show that *SM_ids_pr_mtr_matr* performs better than the rest of Spam Mass variations. We also found that variations of *SM_rds* are meaningful when precision is of immense importance for detection of web spam. In case of web spam detection from entire data set (simply, 100% top), we found *SM_ids_pr_mtr_matr* (when *relativeMass*=0.99) outperforming *SM* up to 1.33 times in terms of recall and 1.02 times in terms of precision. Similarly, *SM_ids* (when *relativeMass*=0.99) outperforms *SM_rds* up to 1.16 times in terms of recall while preserving precision at 0.87. Similarly, for top 10% and 50% *SM_ids_pr_mtr_matr* is performing better than rest i.e., either preserving precision with higher recall or preserving (or comparable) recall with higher precision.

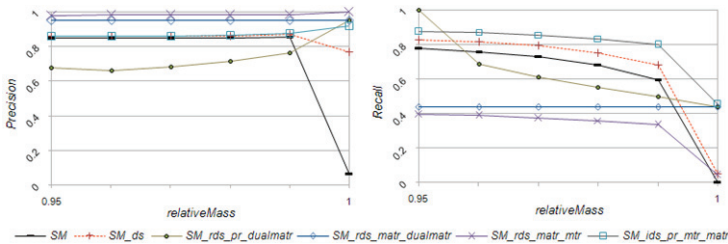


Figure 2. Web spam detection among web nodes falling in the top 10% of PageRank scores

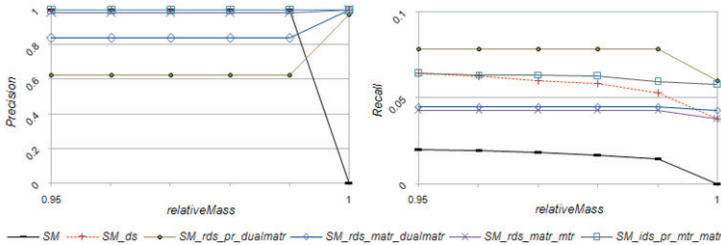


Figure 3. Web spam detection among web nodes falling in the top 50% of PageRank scores

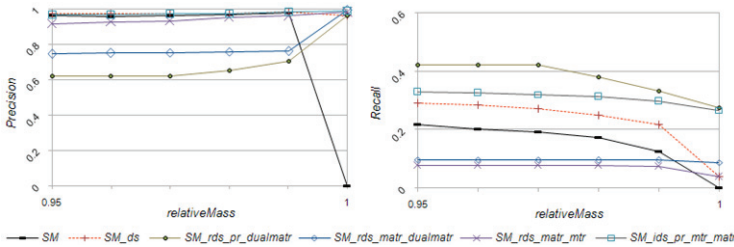


Figure 4. Web spam detection in the entire data set

Table 2. Snapshot of experiments

top	relativeMass	SM		SM_ds		SM_rds_pr_dualmatr		SM_rds_matr_dualmatr		SM_rds_matr_mtr		SM_ids_pr_mtr_matr	
		Pre	Re	Pre	Re	Pre	Re	Pre	Re	Pre	Re	Pre	Re
10%	0.98	1.0	.02	1.0	.06	.62	.08	.84	.04	.98	.04	1.0	.06
	0.99	1.0	.01	1.0	.05	.62	.08	.84	.04	.98	.04	1.0	.06
	1.0	.00	.00	1.0	.04	.97	.06	1.0	.04	1.0	.04	1.0	.06
50%	0.98	.96	.17	.97	.25	.65	.38	.76	.10	.95	.08	.97	.31
	0.99	.98	.13	.98	.22	.71	.33	.76	.10	.96	.07	.98	.30
	1.0	.00	.00	.96	.04	.96	.27	.99	.09	.98	.04	.99	.26
100%	0.98	.85	.68	.86	.75	.72	.55	.95	.44	.98	.36	.87	.83
	0.99	.85	.59	.87	.68	.76	.50	.95	.44	.98	.33	.87	.79
	1.0	.06	.00	.77	.05	.95	.43	.95	.44	1.0	.04	.92	.45

6. CONCLUSION

We propose two new strategies for designing variations of Spam Mass and found both strategies effective in outperforming previous versions of Spam Mass for web spam detection. We found the strategy of inclusive double seeded Spam Mass outperforming original version of Spam Mass by up to 1.33 times in terms of recall and 1.02 times in terms of precision. Similarly,

strategy of inclusive double seeded Spam Mass version outperforms double seeded Spam Mass by up to 1.16 times in terms of recall while preserving precision. Moreover, strategy of reflected double seeded Spam Mass was effective when higher precision is required for the detection of web spam.

REFERENCES

1. **Becchetti, L., Castillo, C., Donato, D., Baeza-YATES, R., and Leonardi, S.** 2008. Link analysis for web spam detection. *ACM Trans. Web*, 2, 1, (Mar. 2008), 1–42.
2. **Brin, S. and Page, L.** 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th conference on World Wide Web (Brisbane, Australia, April 1998)*. WWW'98. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, 107-117.
3. **Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., and Vigna, S.** A reference collection for web spam. *SIGIR Forum*, 40, 2, (Dec. 2006), 11-24.
4. **Duc, P.M., Heo, J., Lee, J., and Whang, K.** 2009. Ranking quality evaluation of PageRank variations. *Journal of the Institute of Electronics Engineers of Korea (in English)*, 46, 5, (Sept. 2009), 14-28.
5. **Gyongyi, Z., Berkhin, P., Garcia-Molina, H., and Pedersen, J.** 2006. Link spam detection based on mass estimation. In *Proceedings of the 32th conference on Very Large Data Bases (Seoul, Korea, September 2006)*. VLDB'06. VLDB Endowment, 439-450.
6. **Gyongyi, Z., Garcia-Molina, H., and Jan, P.** Combating web spam with TrustRank. In *Proceedings of the 30th conference on Very Large Data Bases (Toronto, Canada, August 2004)*. VLDB'04. VLDB Endowment. 576-587.
7. **Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A.S.** 1999. The web as a graph: measurements, models, and methods. In *Proceedings of the 5th annual conference on Computing and Combinatorics (Tokyo, Japan, July 1999)*. COCOON'99. Springer-Verlag, Berlin, Heidelberg, 1-17.
8. **Krishnan, V. and Raj, R.** Web spam detection with Anti-TrustRank. In *Proceedings of the 2nd workshop on Adversarial Information Retrieval on the Web (Washington, USA, August 2006)*. AIRWEB'06. 37-40.
9. **Qureshi, M. A.** 2011 Improving the Quality of Web Spam Filtering by Using Seed Refinement. Master Thesis. KAIST.
10. **Salton, G., and McGill, M.J.** 1983. *Introduction to modern information retrieval*. McGraw-Hill.
11. **Yoshida, Y., Ueda, T., Tashiro, T., Hirate, Y., and Yamana.** 2008. What's going on in search engine rankings?. In *Proceedings of the 22nd Conference on Advanced Information Networking and Applications – Workshops (Okinawa, Japan, March 2008)*. AINAW'08. IEEE Computer Society, Washington, DC, USA, 1199-1204.